Supporting Information

Genomics-Aided Structure Prediction

Joanna I. Sułkowska [1,a], Faruck Morcos [1,a], Martin Weigt [b], Terence Hwa [a], José N. Onuchic [c]

[a]Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093- 0374; [b] Laboratoire de Génomique des Microorganismes, UMR 7238, Universit Pierre et Marie Curie, 15 rue de l'cole de Mdecine, 75006 Paris, France; [c] Center for Theoretical Biological Physics, Rice University, Houston, TX 77005-1827; [1] These authors contributed equally to this work. Correspondence should be addressed: T.H. (hwa@ucsd.edu), J.N.O.(jonuchic@phyics.ucsd.edu).

SI Figures

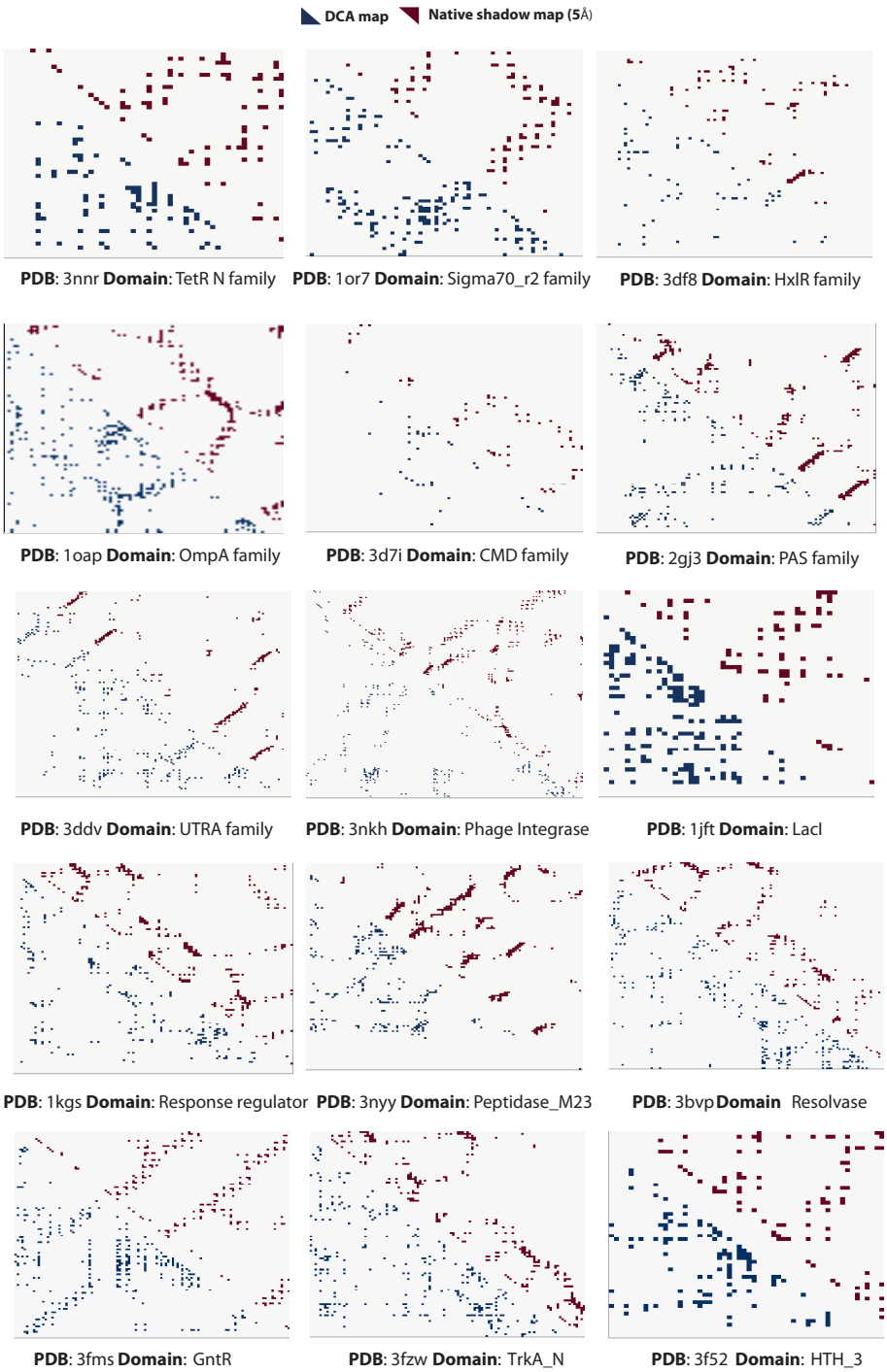

Figure S1: Comparison of estimated contact maps with native maps. Lower triangular maps, below diagonal, represent DCA contact maps and upper triangular maps are native shadow maps with cutoff value of 5Å. All the predictions done in this study used as input a set of contacts estimated using Direct Coupling Analysis (DCA). One of the main observations of the results in the main manuscript is that our methodology is robust to deviations of native contact maps with respect to the estimated ones. DCA produces high quality estimates of contact maps both in terms of true positive predictions but also in terms of the sparsity of the predicted contacts. Other statistical methods like mutual information produce a relatively good number of true positive contacts but they tend to cluster in specific regions that obscure the global structure of the native contact map [1].

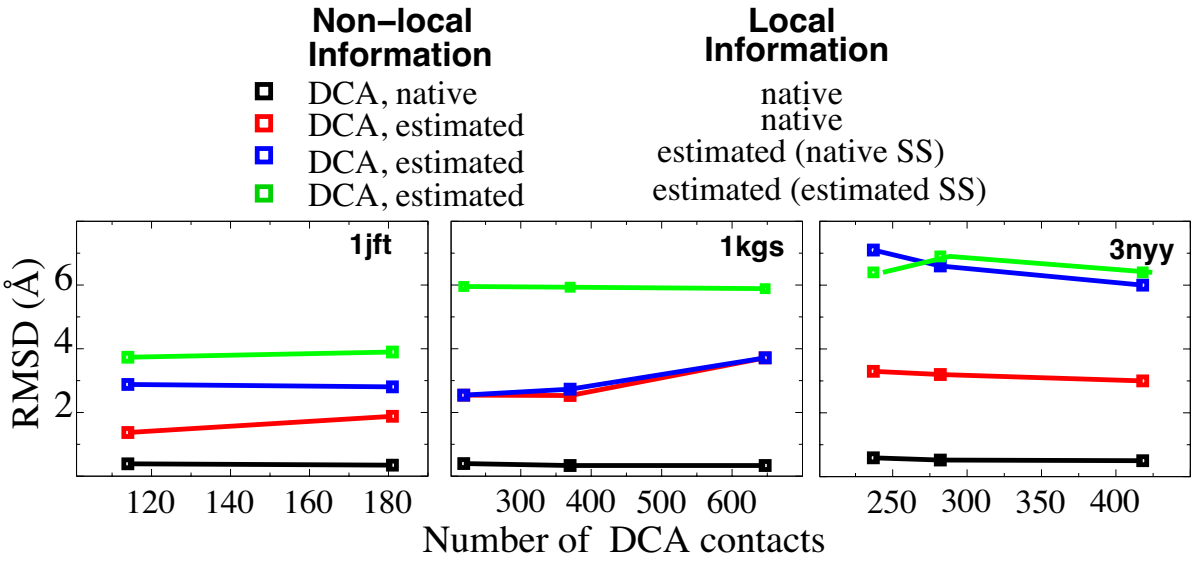

Figure S2: The quality of the predicted structures (testing proteins) measured in RMSD (Å) as a function of different number of DCA contacts.

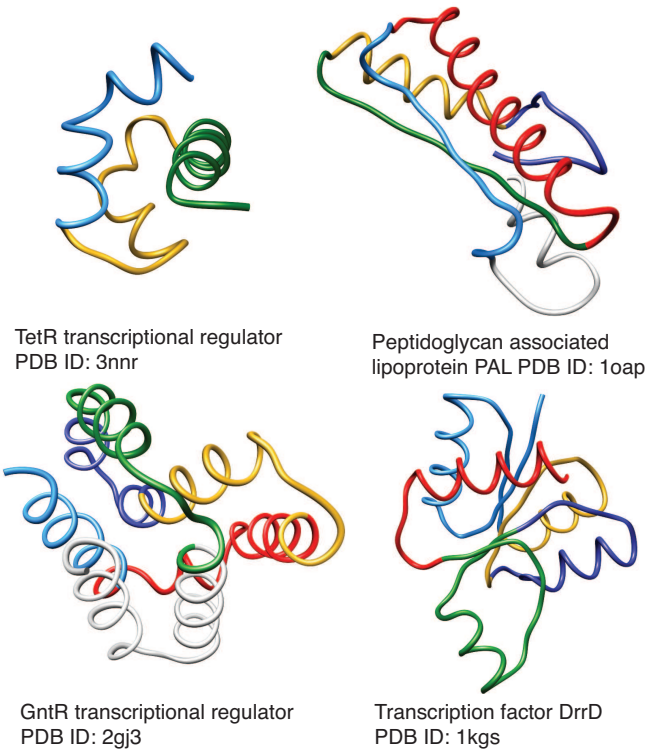

Figure S3: Native structures of exemplary proteins in the main text, Fig. 4. The predictions using contacts from DCA and native knowledge of torsional angles and contact distances in column 1 of Fig. 4 in the main text have a very high resemblance to these native structures. These structures and those in Fig. 4 were produced using UCSF Chimera [4].

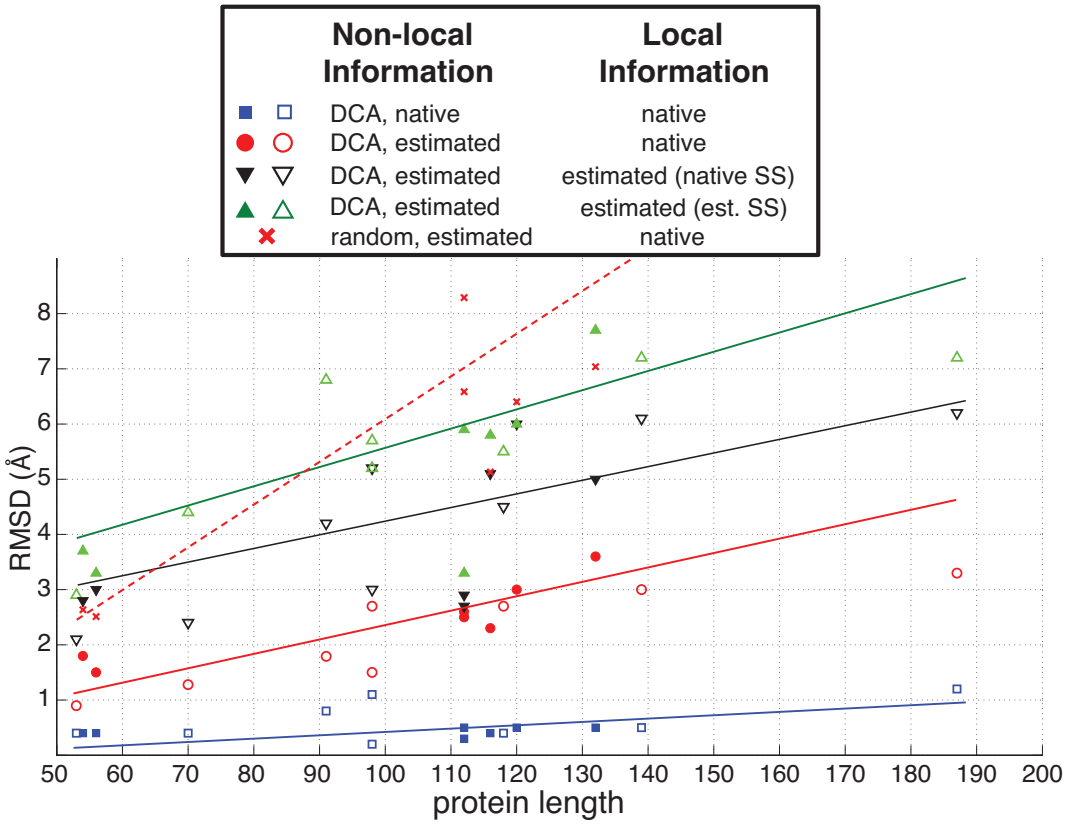

Figure S4: Predicted RMSD for 15 proteins of different sizes for 100% of the protein residues. The symbols indicate the nature of the information on local and non-local residue interactions. Non-local interactions are always derived from DCA contacts. Local information is estimated based on the type of local secondary structure (SS). Open symbols refer to proteins used to derive the statistical potentials, while filled symbols refer to proteins that were used to test this model. The lines are guides to trends by symbols of the same color.

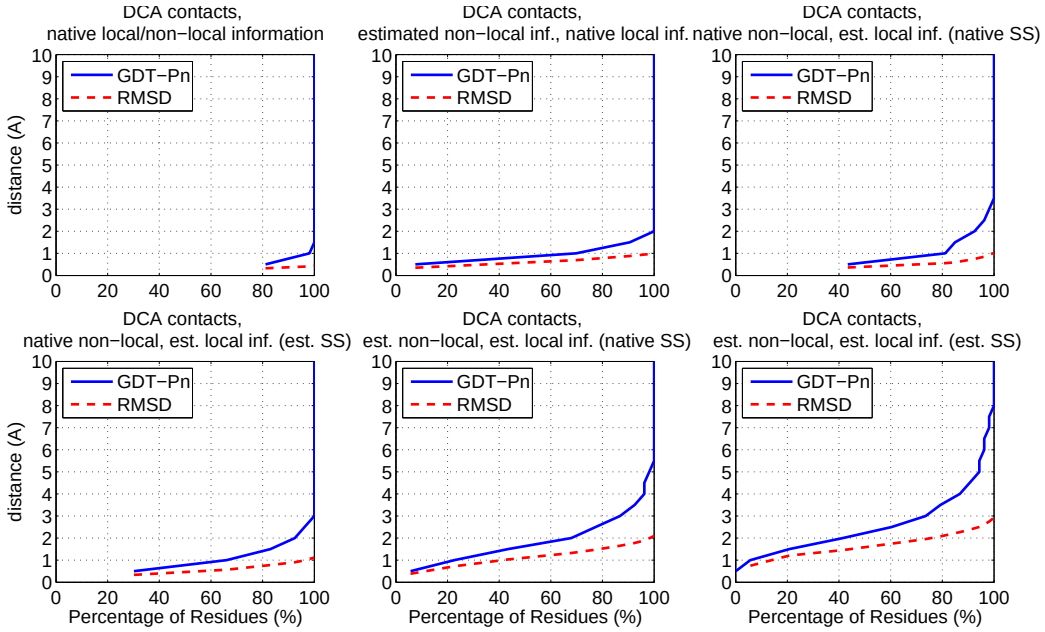

Figure S5: GDT and RMSD curves for the transcriptional regulator of the TetR family, PDB: 3nnr.

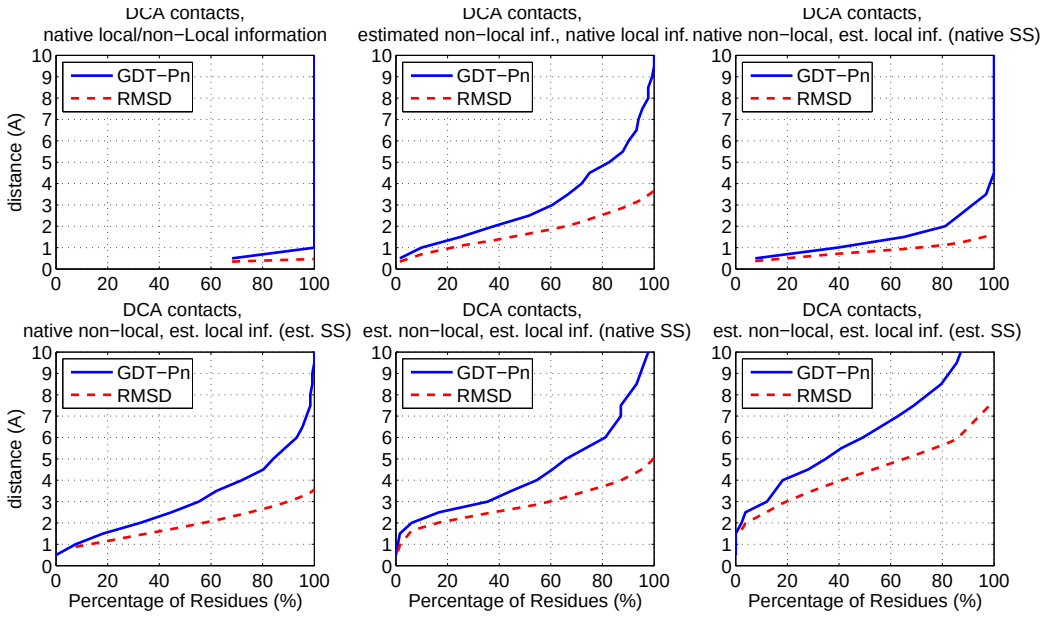

Figure S6: GDT and RMSD curves for the N-terminal catalytic domain of TP901-1 Integrase, PDB: 3bvp. A significant difference in the RMSD values is observed when comparing the top 80% of residues and the complete protein. This is evidence that a small fraction of outliers affect the RMSD for the complete protein.

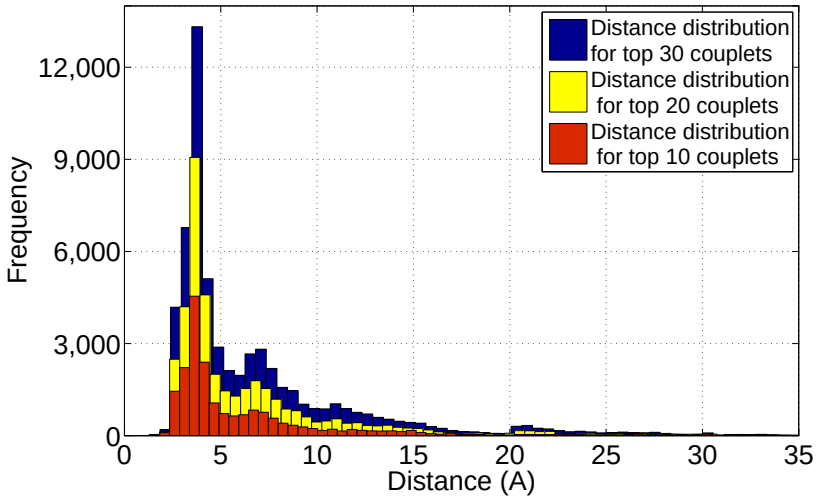

Figure S7: The distribution of minimal atomic distances between the top 10, 20 and 30 predicted residue pairs, using Direct Coupling Analysis method, DCA. A total of 856 PBD structures were used to obtain these statistics.

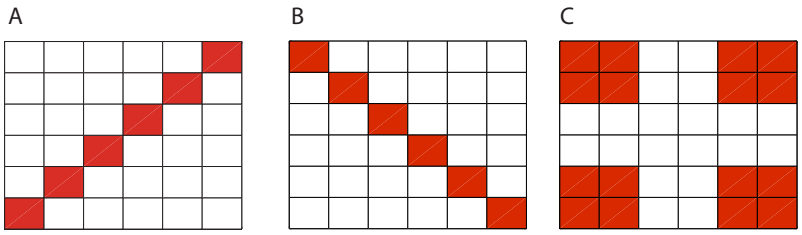

Figure S 8: Contact map patterns used as masks of a 2D convolution for secondary structure identification. A-B) Diagonal patterns used to identify parallel and anti-parallel β-strand/β-strand interactions. C) A chessboard-like pattern is used to identify interactions involving α-helices.

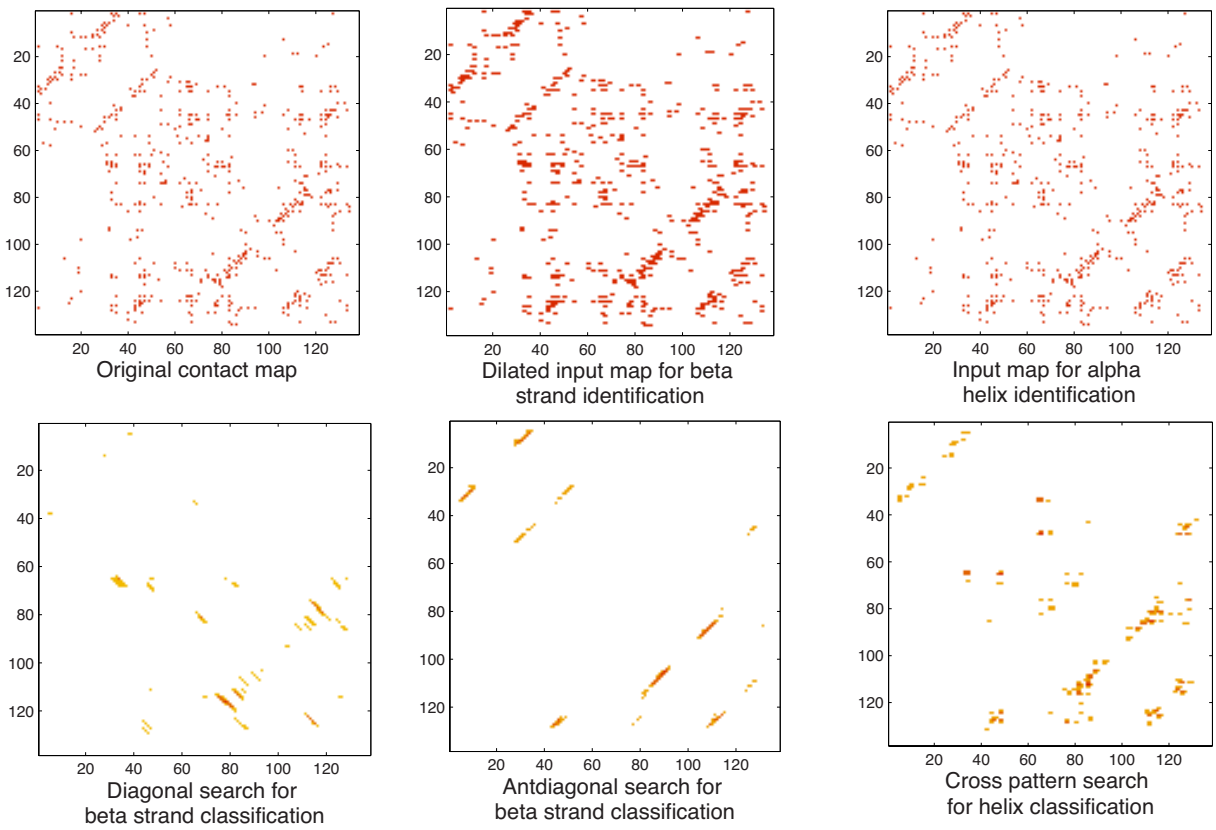

Figure S9: Secondary structure prediction procedure using contact map processing. This figure shows the example of a transcriptional regulator of the GntR family (PDB: 3ddv). The upper left panel shows the estimated DCA contact map. The upper center panel shows a dilated map which is used as pre-processed input for β strand identification. The input for α helix identification does not require pre-processing (upper right panel). Lower panels show the results after applying the 2D convolution masks for diagonal patterns $SS_{\beta_{diag}}$(lower left panel), anti-diagonal patterns $SS_{\beta_{anti}}$ (lower center panel) for β strand identification and inverted cross patterns for α helix classification SS_{α} (lower right panel). These outcomes are used to estimate secondary structure classifications for each residue in the domain based as detailed in Algorithm S1.

SI Tables

Table S1: Annotations for the set of 15 protein domains for which the structural prediction methodology described in the main manuscript was applied. The first 8 proteins were used to determine the proper set of global parameters for the structure based model (SBM).

	Protein (PDB ID)	Fold	Domain Length	Name	Pfam Domain	Organism
Training set	3nnr	α	53	TetR-family transcriptional regulator	TerR_N	*Marinobacter aquaeolei*
	1or7	α	70	RseA	Sigma70 region 2	*Escherichia coli*
	3df8	α	91	Possible HxlR family transcriptional factor	HxlR	*Thermoplasma volcanium*
	1oap	α/β	98	Peptidoglycan associated lipoprotein PAL (Periplasmic domain)	OmpA	*Escherichia coli*
	3d7i	α	98	Oxygen detoxification CMD protein	CMD	*Methanococcus jannaschii*
	2gj3	α/β	118	Transcriptional regulation sensor protein NifL	PAS	*Azotobacter vinelandii*
	3ddv	β	139	Transcriptional regulator GntR family	UTRA	*Enterococcus faecalis*
	3nkh	α	187	Integrase MRSA strain	Phage integrase	*Staphylococcus aureus*
Test set	1jft	α	54	Purine repressor PurR (N-terminal)	LacI	*Escherichia coli*
	3f52	α	57	Gene regulator ClgR	HTH_3	*Corynebacterium glutamicum*
	1kgs	α/β	112	Transcription factor DrrD	Receiver domain (Response regulator)	*Thermotoga maritima*
	3nyy	β	112	Putative glycyl-glycine endopeptidase lytM	Peptidase_M23	*Ruminococcus gnavus*
	3fwz	α/β	116	Inner membrane protein ybaL	TrkA_N	*Escherichia coli*
	3fms	α	120	GntR transcriptional regulator	GntR	*Thermotoga maritima*
	3bvp	α/β	133	N-terminal Catalytic Domain of TP901-1 Integrase	Resolvase	*Lactococcus phage TP901-1*

Table S2: GDT_TS scores for the proteins (100% of residues) involved in this study.

	Non-local Information	Native	Estimated	Native	Native	Estimated	Estimated
	Local Information	Native	Native	Estimated (native SS)	Estimated (estimated SS)	Estimated (native SS)	Estimated (estimated SS)
	PDB ID	GDT_TS					
Training set	3nnr	99.5	92.4	93.4	89.6	71.7	58.5
	1or7	100.0	85.0	92.9	78.9	64.0	44.0
	3df8	94.8	77.2	56.0	34.6	45.3	26.4
	1oap	100.0	81.4	76.0	65.3	55.6	37.2
	3d7i	97.7	56.6	69.1	51.5	37.5	34.9
	2gj3	99.6	62.0	75.6	72.0	40.8	36.3
	3ddv	99.1	57.9	68.7	51.0	30.0	24.1
	3nkh	84.8	51.7	54.0	61.4	34.8	26.0
Test set	1jft	100.0	80.6	76.0	76.4	58.8	47.7
	3f52	99.6	78.1	90.6	83.9	53.1	51.3
	1kgs	100.0	64.4	94.2	73.8	61.3	32.7
	3nyy	98.4	53.6	67.2	59.6	33.5	27.0
	3fwz	99.4	66.4	90.7	66.4	37.5	33.8
	3fms	100.0	62.0	94.7	92.2	58.0	52.7
	3bvp	100.0	54.3	79.7	52.7	37.8	23.6

Table S3: Comparison of performance of predicted protein structures with respect to experimentally determined structures. The parameters are: contact maps based on DCA, random maps and estimated non-local information.

Contact maps		**DCA**	**Random**
Non-local Information		**Estimated**	**Estimated**
Local Information		**Native**	**Native**
Protein ID (fold)	**Length/No. of contacts**	**RMSD in Å(RMSD 100% of residues)**	
1jft(α)	54/114	1.2 (1.4)	2.1 (2.6)
3f52(α)	57/111	1.2 (1.5)	2.0 (2.5)
1kgs(α/β)	112/219	1.8 (2.5)	4.9 (6.5)
3nyy(β)	112/237	2.5 (3.0)	6.8 (8.2)
3fwz(α/β)	116/271	1.7 (2.3)	4.0 (5.1)
3fms(α)	120/301	2.0 (2.6)	5.3 (6.7)
3bvp(α/β)	133/301	2.5 (3.6)	4.6 (6.2)

Table S4: RMSD of sample proteins (100% of residues) from Table 1 before/after refinement.

Non-local Information		**Native**	**Estimated**	**Estimated**	**Estimated**
Local Information		**Native**	**Native**	**Estimated (native SS)**	**Estimated (estimated SS)**
Protein ID	**Length**	**RMSD in Å**			
3nnr	53	0.4	0.9/1.0	2.1/2.4	2.9/2.3
1oap	98	0.2	1.5/1.6	3.0/3.5	5.1/4.9
2gj3	118	0.4	2.7/2.9	4.3/4.2	5.0/5.4
1kgs	112	0.3	2.5/2.5	4.3/4.1	5.4/5.9

Table S5: The best prediction performance based on optimized parameters for the statistical distance potential (non-local information) and the same number of DCA contacts as in Table 1 (100% of residues) in the main manuscript. Results shown in Å.

Non-local Information	**Native**	**Estimated**	**Estimated**	**Estimated**
Local Information	**Native**	**Native**	**Estimated (native SS)**	**Estimated (estimated SS)**
Protein ID (fold)	**RMSD in Å**			
3nnr(α)	0.3	0.1	2.1	2.9
1or7(α)	0.3	0.9	2.1	4.1
3df8(α)	0.6	1.1	3.9	6.4
1oap(α/β)	0.2	1.5	3.0	5.0
3d7i(α)	0.9	2.7	5.2	5.6
2gj3(α/β)	0.4	2.7	4.3	5.4
3ddv(β)	0.4	2.7	6.1	7.2
3nkh(α)	0.9	3.3	6.2	7.2

Table S6: Best predictions (100% of residues), based on different number of DCA contacts. Results shown in Å. For the estimated local information, SS represents the knowledge of native secondary structure while $\widehat{SS}$ represents estimated secondary structure information.

Non-local Information		**Native**		**Est.**		**Native**		**Native**		**Est.**		**Est.**
Local Information		**Native**		**Native**		**Est.-SS**		**Est.-$\widehat{SS}$**		**Est.-SS**		**Est.-$\widehat{SS}$**
PDB ID (fold)	**DCA contacts**	**RMSD in Å**	**DCA contacts**	**RMSD in Å**	**DCA contacts**	**RMSD in Å**	**DCA contacts**	**RMSD in Å**	**DCA contacts**	**RMSD in Å**	**DCA contacts**	**RMSD in Å**
3nnr(α)	178	0.3	178	0.5	137	1.0	137	1.1	83	2.0	83	2.8
1or7(α)	292	0.3	483	0.9	165	1.0	165	2.3	203	2.1	203	4.1
3df8(α)	93	0.6	93	1.6	62	3.0	52	5.2	62	3.9	62	6.4
1oap(α/β)	638	0.3	423	0.9	1087	0.6	1087	1.0	104	3.0	282	5.0
3d7i(α)	510	0.6	183	1.7	76	2.3	76	3.4	50	5.2	101	5.7
2gj3(α/β)	367	0.4	367	2.7	367	2.0	367	2.8	367	4.3	367	5.5
3ddv(β)	264	0.5	264	2.2	264	2.3	298	3.9	298	5.6	298	6.8
3nkh(α)	1254	0.5	445	3.3	726	3.0	726	3.3	455	6.2	445	7.2

SI Methods

1 Direct Coupling Analysis

Direct Coupling Analysis (DCA) The mathematical formulation and an extensive evaluation of the capabilities of the *mean field* implementation of DCA (mfDCA) is described in detail by Morcos et al. [1].

1.1 Multiple Sequence Alignments (MSA)

The multiple sequence alignments used in our study were obtained directly from Pfam for the protein families listed in Table S1. The same alignment parameters as the ones used in the Pfam database were employed. After obtaining the full alignment from Pfam, a post-processing step was needed to remove all the inserts (represented by lowercase amino acids) and the gaps introduced to align those inserts (periods). That way we only retain residues mapped to the Hidden Markov Model of the protein family. No other parameter optimization was done to the MSA. The main purpose in doing this was to keep the number of adjustable parameters at a minimum.

1.2 Distance distribution

One important feature of DCA is the quality of its contact predictions. Since DCA tries to disentangle direct from indirect correlations among residue pairs, the resulting highly ranked predictions tend to be residues that are in physical proximity in the three dimensional structure of a given protein sequence [1]. The distribution of minimal atomic distances between the top 10, 20 and 30 predicted residue pairs, using DCA, is illustrated in Figure S7. These statistics were computed for 856 PDB structures for which residue-residue prediction was performed. It is clear that the peak of the distribution is found between 3.5-5Å with a second peak around 7-8Å.

2 Structure-Based Model is parametrized by the native state

We used a structure based model, where each amino acid is represented by a single bead of unit mass placed at the location of the C_α atom. Bond lengths are maintained by harmonic potentials. Non-bonded atom pairs, that are in contact in the native state between residues i and j (where $|j-i|>4$), are given an attractive Gaussian well potential [8]. All other non-local interactions are repulsive. The amino acids a_i and a_j that are in contact in the native state are identified based on Shadow map [7]. The basic form of the potential is,

$$V(r_{ij}) \quad = \quad V_{contact}(r_{ij}) + V_{tor}(\alpha_i, \tau_i). \tag{1}$$

The contact potential is composed of two terms:

$$\begin{aligned} V_{contact}(r_{ij}) \quad = \quad & \sum_{\substack{\text{DCAcontacts} \\ \text{(i,j>i+4)}}} \epsilon_C \left[(1+(\sigma^C/r_{ij})^{12})(1-\exp(-(r_{ij}-r_{ij}^N)^2/(2(\sigma_{ij}^N)^2)))-1\right] + \\ & \sum_{\substack{\text{non contacts} \\ (i,j>i+4)}} \epsilon_R(\frac{\sigma^C}{r_{ij}})^{12} + \sum_{\text{bonds}} k_b(r_{ij}-r_{ij}^{Nb})^2, \end{aligned} \tag{2}$$

where $V_{contact}(r_{ij})$ is a Gaussian well. r_{ij}^N corresponds to the native distance between the pair i,j and width σ_{ij}. When contact maps and distances were determined based on crystallographic data, σ_{ij} is defined such that $V_{ij}(r_{ij}^N = 1.2\mu_{ij}) = -\frac{1}{2}$, which gives the Gaussian well a variable width that mimics the width of a 10-12 Lennard-Jones interaction with the same r_{ij}^C. This choice defines $(\sigma_{ij}^N)^2 = \left(r_{ij}^N\right)^2/(50\ln 2)$. The second term is independent of r_{ij}^N and maintains the excluded volume of the polypeptide. The parameter $\sigma^C = 4$Å corresponds to the repulsive size of the beads, between both native and nonnative pairs. The last term represent interaction between beads adjacent in the sequence separated by native distance, r_{ij}^{bN}.

The local propensity of the chain is described by a traditional dihedral potential,

$$V_{tor}(\alpha_i,\tau_i) = \sum_{\text{angle\{i\}}} k_a(\tau_i-\tau_i^N)^2 + \sum_{\text{dihedral\{i\}}} k_d([1-\cos(\alpha_i-\alpha_i^N)] + \frac{1}{2}[1-\cos(3(\alpha_i-\alpha_i^N))]) \quad (3)$$

where τ_i^N is the native angle between the bonds connecting three consecutive atoms, and α_i^N is the native angle between the planes defined by four consecutive atoms. The interactions strengths are $k_b = 2\times10^4\epsilon/nm^2$, $k_a = 40\epsilon/rad^2$ and k_d= ϵ, with the reduced unit of energy $\epsilon = k_BT$. Symbol N is used to refer to a single native structure as the reference state. This model has been characterized in detail elsewhere [9] and is freely available on the web [7].

3 Structure based model combined with statistical potentials

The prediction of protein structures was performed based on the energy function given by equation 1, where all parameters with superscript N and the shape of the Gaussian basin were replaced with values obtained from statistical potentials. The symbol r_{ij}^{est} represents the estimated distance between the pair i,j with correspoding σ_{ij}^{est}. The native distance between consecutive atoms that are in contact along sequence r_{ij}^{bN} is replaced by the constant value r_{ij}^b =3.8Å.

3.1 Contact maps

The native contact maps were replaced with maps obtained based on Direct Coupling Analysis (DCA). The DCA maps were sometimes modified. Some pairs were removed from the map when they had very small probability of existence based on methods described in the following section.

In order to decide the number of DCA contacts to use as input to DCA-fold, we systematically tested different numbers of DCA contacts for each of the training proteins until we found the optimum prediction. For the testing proteins, we used similar number of DCA contacts as the ones observed in the training set, based on the protein which have to most similar number of amino acids. In general, we observe that the prediction results are robust to the specific number of DCA contacts selected (See Figure S2).

In order to decide the number of DCA contacts to use as input to DCA-fold, we test the quality of the prediction (RMSD) based on at least 4 different set of DCA contacts.

3.2 Statistical potential for non-bonded interactions

We developed a series of distance potentials based on [15, 14, 16, 17] to mimic the interaction between pairs from DCA map. These potentials have a minimum at the estimated pairwise distances r_{a_i,a_j}^c, parameters which weight different type of interactions, chemical properties of the amino acids types a_i and a_j and their sequence separation S_{ij}. We optimized the coefficients in our potential in a way to obtain a minimally frustrated landscape based on training set composed of 8 proteins.

Below we describe the main steps to construct a knowledge based distance potential, to model the type of interactions between DCA pairs. The following steps are optimized independent of the testing set data.

1. As an input to construct the distribution of pairwise distances r_{a_i,a_j}^t, sets of proteins were used e.g.:
   a test set of 32 proteins (S_1) from [16], a test set of 65 proteins (S_2) from [19], a test set of 60 proteins (S_3) from [18].

2. To detect pairs of interacting amino acids a_i and a_j in the native state of proteins and their distances r_{a_i,a_j}^t in each test set $S_{1,2,3}$, we used two types of techniques: shadow map [7] and cutoff map [20].

3. When all pairs a_i,a_j with associated distance r_{a_i,a_j}^t were detected, they have been grouped based on the type of chemical properties K_{a_i,a_j}^{xy}. Where x and y represent four possible types of amino acids properties, hydrophilic (h), polar (p), acid/hydrophobic (a) and basic (b).

4. For each amino acid pair a_i,a_j, the distribution of sequence separation, S_{ij}=$|j-i|$ was calculated.

5. The probabilities for each amino acid to be in a particular pairwise conformation were computed based on the three most probable sets as follows:

$$P^{hh}(a)=(N^a_{hh}/N_{hh})/(N^a/N),$$
$$P^{bb}(a)=(N^a_{bb}/N_{bb})/(N^a/N),$$
$$P^{hb}(a)=(N^a_{hb}/N_{hb})/(N^a/N).$$

Where N^a_{xy} is the total number of pairs containing amino acid a involved in a pairing with chemical properties x, y. On the other hand, N^a is the frequency of cases where amino acid a appears in the data set. N_{xy} is the total number of pairs with chemical properties xy and N is the total number of amino acids found in the dataset. These values were used to obtain the probability $P^{xy}(a_i, a_j) = -\ln P^{xy}(a_i) + \ln P^{xy}(a_j)$.

In the following, we use information from our previous steps to determine $r^{est}_{a_i,a_j}$ (based on distribution of $\mathrm{r}^t_{a_i,a_j}$) and the shape of the Gaussian basin for each pair of amino acids predicted with DCA for a given protein in the training set.

1. The distribution of pairwise distances $\mathrm{r}^t_{a_i,a_j}$ with associated chemical property $K^{x,y}_{a,a}$ and sequence distance S_{ij} can be broad or have more than two peaks. To find what is the most probable value of the pairwise distance $r^{est}_{a_i,a_j}$ for a given pair from DCA map, we use the ranking of DCA contacts as defined in Section 1. Contacts from the DCA map are divided into three sets. The following possible combinations of sets were tested: the first 10 contacts and the rest of the contacts, the first 20 contacts and the rest of the contacts, the first 30 contacts and the rest of the contacts. Depending on which data set the pair is associated, $r^{est}_{a_i,a_j}$ is assigned a shorter (either top 10, 20 or 30) or longer distance (rest of the contacts) based on the distribution of pairwise distances $\mathrm{r}^t_{a_i,a_j}$.

2. All pairs are then grouped according to their sequence proximity classes (short, intermediate and long range), similarly to an associative memory model [16]. We distinguish the following subgroups: short range in sequence contacts $|j-i| <= 4, 5$; intermediate range in sequence contacts $5 < |j-i| <= 8$, $5 < |j-i| <= 10$, $5 < |j-i| <= 13$, and long range in sequence contacts $|j-i| > 8$, 10, 13.

3. According to which group from the previous step a given pair was assigned, the optimal shape of the Gaussian potential (given by Eq. 2) was tested. The parameter ϵ, which describes the strength of interaction, was used always as $\epsilon = 1$. We optimized width $\sigma^{est}_{a_i,a_j}$ of the basin. Both of these parameters, independently, vary the shape of the attractive potential without changing its repulsive part. The following potential widths were tested:

   (a) $\sigma^{scaled}_{a_i,a_j}$, the scaling behavior of the shape of the potential was modeled by making the width proportional to the contact distance, σ^{scaled}_{ij}=$kr^{est}_{a_i,a_j}$, where the constant k was chosen to be (k = 0.091).

   (b) $\sigma^{short,medium,long}_{a_i,a_j}$ was represented by three different constant widths:

   - For short range contacts distance (around 5.5Å), typically assigned to hydrogen bonds and contacts that were predicted with high confidence by DCA method, we used $\sigma^{short}_{a_i,a_j}$= 0.4, 0.5, 0.6 Å.
   - For intermediate range contact distances (around 7-8Å), we used $\sigma^{medium}_{a_i,a_j}$= 0.7 Å.
   - For long range contacts distances (above 10Å), typically hydrophobic interactions and contacts which were not predicted with high confidence, we used $\sigma^{long}_{a_i,a_j}$= 1.0, 1.2, 1.5 Å. Such broad width of the well of the Gaussian potential was still maintaining ability to fold, however folding was less cooperative.

4 Torsional potential

The native geometry of the amino acid chain, the native angle τ^N_i and the native dihedrals α^N_i, were replaced by τ^{est}_i and α^{est}_i based on the following procedures.

4.1 Conversion from all atom to Cα representation for dihedral angles

The degrees of freedom of peptide bonds are usually described with ϕ and ψ torsional angles (dihedral angles) based on N-C$_\alpha$ and C$_\alpha$-C bonds, respectively. To describe the conformation of the backbone in C$_\alpha$ model we used a relation developed by Levitt [12]. The degrees of freedom of the backbone are given by two angles α and τ, which are obtained from the following relation:

$$\alpha_i = 180^\circ + \phi_{i+1} + \psi_i + 20^\circ(\sin\phi_i + \sin\psi_{i+1}) \quad (4)$$

based on [12]. The torsional angle α_i is defined by the position of four adjacent C_α atoms, C_{i-1}, C_i, C_{i+1}, C_{i+2}, and two pairs of $\phi_{i-1,i}$ and $\psi_{i-1,i}$ values. The bond angle τ_i is defined between C_{i-1}, C_i and C_{i+1} and changes with respect to α_i as:

$$\tau_i = 106^\circ + 13^\circ \cos(\alpha_i - 45^\circ). \tag{5}$$

4.2 Dihedral angle estimation

To estimate dihedral angles based on protein sequences we used neighbor dependent probability distributions calculated by Ting el al. [5]. This pairwise Ramachandran distributions can be combined to get estimates of ϕ and ψ angles using the following formulation:

$$(\hat{\phi}, \hat{\psi}) = \arg\max \hat{p}(\phi, \psi | C, L, R) \tag{6}$$

with

$$\hat{p}(\phi, \psi | C, L, R) = \frac{\hat{f}(\phi, \psi | C, R)\hat{f}(\phi, \psi | C, L)}{\hat{f}(\phi, \psi | C)} \tag{7}$$

where C,L,R are the center, left and right position in the sequence and $\hat{f(\cdot)}$ is the statistical pairwise distribution. Results of Eq. 6 are in general biased towards alpha helix prediction. To correct for that bias, we used knowledge of secondary structure (either native or estimated) to guide dihedral estimation. If it is known *a priori* that a given sequence triplet belongs to a β-strand or α-helix then we restrict the estimates to a preferred quadrant of typical Ramachandran distributions for β-strands or α-helices. This way the maximization procedure in Eq. 6 would get the angles with highest probability constrained to such predefined quadrants. For the case of the alpha helix we constrain the estimates to a region where $\hat{\phi} < -60$ and $-90 < \hat{\psi} < -40$. We defined the corresponding region for beta strands as being $\hat{\phi} < -100$ and $\hat{\psi} > 80$. For the rest of possible configurations, like loops, turns, left handed alpha helices, etc. we do not constrain the estimation and use the plain formulation shown in Eq. 7. When we use the native knowledge of the SS classification to bias equation 6 and to estimate (α_i, τ_i), we refer to this with the identifier for the Torsional Potential: **Estimated (native SS)** in the main text and tables.

4.3 Secondary structure estimation

As described in the previous section, we use information about the secondary structure of the protein or domain to guide the estimation of dihedral angles. We can use native knowledge of the secondary structure, i.e. a mapping between a residue and a coarse grained secondary structure category like alpha helix, beta strand and everything else directly from the three dimensional structure. The second option is to infer secondary structure categories using a statistical estimation method. Although there are a number of methods for secondary structure prediction which can be used as input to our model, we decided to extract this information from the DCA contact maps. We use this simple method as a lower limit of what can be achieved with secondary structure prediction, while our upper limit is the native knowledge extracted from the structure.

To determine if a given residue belongs to a secondary structure classification, we search in the contact map for features that could represent secondary structure elements. For example, parallel or anti-parallel aligned β-strands usually form diagonal patterns (Figure S8A-B) and contacts involving α-helices are characterized by a series of chessboard like pattern or inverted cross (Figure S8C). Since the estimated contact maps are usually sparse, we pre-process the map with an image dilation technique that will help merge isolated regions of the contact map that were empty. This is important for contact maps that have a small number of predicted contacts. After this, pattern matching is accomplished using a two dimensional convolution between the estimated contact map and a mask having a diagonal pattern (β-strands) and an inverted cross pattern (α-helices). The output of the convolution will show higher signals in the contact map where the pattern matches better. The output of the beta strand convolution is then compared with the one of the alpha helix convolution. Using a predefined threshold optimized with native contact maps, the algorithm provides a secondary structure assignation (beta, alpha or other) for each residue in the protein. The secondary structure prediction algorithm is summarized in Algorithm S1. This method is used to guide the estimates in Eq. 6 and to determine $(\alpha_i^{est}, \tau_i^{est})$. These values are used as parameters for the torsional potential $V_{tor}(\alpha_i, \tau_i)$. This is referred in the main text and tables as **Estimated (estimated SS)**.

Algorithm S1. Secondary Structure Prediction Algorithm

Input: DCA contact maps: CM

$\mathrm{SS}_\alpha \leftarrow \mathrm{conv2D}(CM, \mathrm{mask}_\alpha)$ ▷ 2D convolution: α-mask
$\mathrm{SS}_\beta{}^{diag} \leftarrow \mathrm{conv2D}(CM, \mathrm{mask}_{\mathrm{diag}})$ ▷ 2D convolution: diagonal β-mask
$\mathrm{SS}_\beta{}^{anti} \leftarrow \mathrm{conv2D}(CM, \mathrm{mask}_{\mathrm{anti}})$ ▷ 2D convolution: anti-diagonal β-mask
$\mathrm{SS}_{\mathrm{ij}} \leftarrow \mathrm{SS}_{\beta_{\mathrm{diag}}} + \mathrm{SS}_{\beta_{\mathrm{anti}}} - 2\mathrm{SS}_\alpha$ ▷ Normalize and compare β and α elements
$SS \leftarrow \sum_j SS_{ij}$ ▷ Collapse the matrix $\mathrm{SS}_{\mathrm{ij}}$ into a one dimensional array SS of length L
▷ where L is the protein sequence length

for All i residues in SS **do**
  if $SS_i > T$ **then** ▷ T is optimized with native contact maps
    $\widehat{SS}_i \leftarrow \beta-$strand
  else if $SS_i < -T$ **then**
    $\widehat{SS}_i \leftarrow \alpha-$helix
  else
    $\widehat{SS}_i \leftarrow$ other ▷ this could be loops, turns and left-handed helices
  end if
end for
Output: Secondary structure prediction $\widehat{SS}$

The secondary structure assignations in the estimate $\widehat{SS}$ are used to guide the torsional angle estimates $(\alpha_i^{est}, \tau_i^{est})$ discussed in the previous section. Figure S9 shows an example of the processing done to predicted DCA contact maps to identify secondary structures.

4.3.1 Secondary structure prediction metric

The performance metric used for our DCA contact map based predictor of secondary structure (SS) is a Hamming distance between the secondary structure classifications (beta strands = 2, alpha helices=1, and other possible configurations=0) and the predictions. Suppose that we have an native **SS** vector in $\mathbb{F}^L$ for a given protein of length L and an alphabet $\mathbb{F} \in \{0, 1, 2\}$ and an estimated vector $\widehat{\mathbf{SS}}$ in $\mathbb{F}^L$. Then the performance criterion will try to minimize the Hamming distance between the vectors defined as:

$$H(\mathbf{SS}, \widehat{\mathbf{SS}}) = \sum_{i=1}^{L} SS_i \wedge \widehat{SS}_i \tag{8}$$

where the $\wedge$ operator only equals 1 when the the elements are the same and otherwise is 0. In case of a tie in the minimum distance, then estimates of $SS_i = 0$ will have a better priority since these only represent unguided dihedral estimates from the Drichlet distributions. Other optimizations can be done if it is known *a priori* if a protein is only alpha helical or contains purely beta strands. However, such considerations are not used in this work.

5 Best parameters for statistical potentials to predict protein structures

The energy function constructed according to the above criteria was next optimized based on a different number of DCA contacts for eight proteins from the training set (Table S1). We found that the best results we obtained were those for the model constructed based on :

- Set of 65 proteins from [19] (set S_2) was chosen to build the pairwise distance distribution $r^t_{a_i,a_j}$.
- Pairs of contacts with corresponding distance $r^t_{a_i,a_j}$ were identified with shadow map.
- Pairs with $r^t_{a_i,a_j}$ were clustered according to their chemical properties.

- Pairs from DCA maps were accepted when $P^{xy}(a_i, a_j) = -\ln P_{hh}(a_i) + \ln P_{hh}(a_j)$ was bigger than 0.2.
- The optimal value of pairwise distance $r^{est}_{a_i,a_j}$ was chosen from $r^{t}_{a_i,a_j}$ with the following rule: if a pair was found in the top 20 DCA ranking and the distance distribution $r^{t}_{a_i,a_j}$ had more than one maximum, we chose the maximum corresponding to the shorter distance in the distribution of $r^{t}_{a_i,a_j}$, according to the distribution of distances shown in Figure S7.
- Next, optimal values of $r^{est}_{a_i,a_j}$ were clustered into subgroups $|j-i| =< 5$, $5 < |j-i| =< 8$, $8 < |j-i| <= 12$ and $|j-i| > 13$ to which the corresponding shape of the Gaussian well was assigned: $|j-i| =< 5$ with $\sigma^{fixed,est}$=0.5, $5 < |j-i| =< 8$ with $\sigma^{fixed,est}$=0.7, $8 < |j-i| <= 12$ with $\sigma^{fixed,est}$=1.0, $|j-i| > 13$ with $\sigma^{fixed,est}$=1.2.
- The final results for all fixed parameters considering the previous steps and the method to calculate torsional angles (α,τ) are shown in Table 1 in the main manuscript.

5.1 Molecular dynamics simulations of folding

The simulations were performed with the GROMACS 4.0.5 software package [11]. The software was additionally modified to support different contact and torsional potentials. Reduced units were used for all calculations with time steps of size 0.0005. Trajectory coordinates were saved every 100 time steps.

We performed stochastic dynamics with annealing protocol and the Nose-Hoover thermostat [10] . The annealing protocol was specified as a single sequence of corresponding time steps and reference temperatures. T_f was determined for three proteins and averaged to determine a T^*_f for the rest of the proteins. We used a high temperature of T=160 which corresponds to T=1.4T^*_f (strongly favoring the unfolded configuration) and a few very low temperatures T= 0.16T^*_f, 0.2T^*_f, 0.33T^*_f (strongly favouring the folded configuration).

Annealing starts at T= 1.4T^*_f at 0 ps and stays constant until 20×10^6 time steps. Subsequently, temperature will drop linearly to reach T=0.33 T^*_f at 40×10^6 time steps, and then stays constant until 50×10^6 time steps. The time steps were reduced from 40×10^6 to 20×10^6 time steps and from 50×10^6 to 40×10^6 time steps for the smallest proteins. Initial conformations were generated by simulating homopolypeptides without any attractive interactions. Additionally the long equilibration (1×10^6 time steps) at high temperature was used to ensure an uncorrelated random distribution of starting configurations. When knowledge-based potentials, described in previous sections, were used, they were introduced at the start of the annealing procedure. For each protein, at least 300 independent runs were performed. As minimization is inherently statistical, this ensures convergence on the resolution of protein structures.

6 All-atom reconstruction and empirical all-atom force-field for refinement

To reconstruct all heavy atoms in predicted protein conformations we used PULCHRA software [13]. The following additional PULCHRA options were used, optimization of backbone hydrogen patterns and detection of cis-proline conformations. After reconstruction, the predicted structures were additionally relaxed with an empirical all-atom force-field for refinement. We used Amber99 (as a force field in GROMACS) with explicit Tip3p solvent and counter ions [21]. We used stochastic dynamics with a time step of 2 fs, and Particle Mesh Ewald electrostatics [22].

The RMSD for 4 sample proteins after refinement is shown in Table S4. We found that refinement did not change the fold of the predicted proteins, the RMSD along C_α atoms vary no more than 1 Å with respect to the native conformation. This shows that the initial estimates did not have steric clashes or unphysical conformations.

7 Global Performance Measures

7.1 Q-metric

Another metric for protein structure comparison is the Q-metric [2]. This metric is independent of alignment and compares internal distances of the reference protein structure with the internal distances of a target. Internal distances are calculated between C_α atoms of each residue i and all $N-1$ other C_α atoms in the protein. For a residue pair i, j, it is defined as:

$$Q_{ij} = \exp[-(r_{ij} - r^N_{ij})^2] \tag{9}$$

where matrix r_{ij}^N contains all the internal distances of a given structure for $i > j$. To have a global metric for structure prediction, an average of all Q_{ij} is computed, yielding

$$Q_{total} = \langle Q_{ij} \rangle. \tag{10}$$

7.1.1 Global Distance Test (GDT) metric

Another measure useful for structure prediction comparison is the Global Distance Test (GDT). This metric computes the percentage of residue distances, from the predicted structure with respect to the target structure, under a certain threshold [2, 3]. For instance GDT_Pn represents this percentage under a distance threshold less or equal than nÅ. The GDT total score or GDT_TS is defined as:

$$\text{GDT\_TS} = (\text{GDT\_P1} + \text{GDT\_P2} + \text{GDT\_P4} + \text{GDT\_P8})/4. \tag{11}$$

Performance curves can be calculated for any user defined cutoff n. Figure S5 and Figure S6 show GDT curves (blue) for a two proteins (PDB: 3nnr, 3bvp). RMSD as a function of a fraction of the total number of protein residues (red) is also shown. A significant difference in the RMSD values is observed when comparing the top 80% of residues and the complete proteins. Illustrating how a small fraction of outliers has large effects on the RMSD for the complete protein. Table S2 shows the GDT_TS scores for the rest of the proteins in the dataset for difference amounts of native information used in the prediction.

8 Random maps

To evaluate the contribution of the DCA contacts, we generated 7 random contacts maps with the same number of contacts as the DCA estimates for each protein from the testing set. Those maps included only contacts with $|j-i| > 4$ as DCA maps. We performed prediction for each protein using these maps, estimated non-local information and native local information. The median RMSD value obtained from the best predicted structures based on 7 random maps and its 80% is shown in Table S3.

9 Protein sets and summary of results

The model obtained based on the training set data (8 proteins) represents the energy function which successfully predicted protein structures, see Table 1 in the main manuscript. We used this best general model with fixed parameters to predict the structures of 7 proteins domains ranging in length from 50 to 135 residues (lower part of Table 2).

We found that a large variation of the number of DCA contacts has an influence on the quality of the prediction (Table S6), as should be expected. However two trends are observed: firstly, increasing the number of DCA contacts, even with more false positives, improved resolution of prediction when local information is known. Secondly, when local information is estimated, a better prediction is observed with a smaller number of DCA contacts.

A comparison between Table 1 in the main manuscript, showing results with fixed model parameters, and Table S5, with individual parameter optimizations, shows small performance fluctuations (gains up to 2Å). This indicates the robustness of our designed model.

SI References

References

[1] Morcos F et al. (2011) Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* 108: E1293–E1301.

[2] Ben-David M et al. (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins: Struct Func and Bioinf* 77: 50–65.

[3] Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acid Res* 31(13): 3370–74.

[4] Pettersen EF et al. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

[5] Ting D et al. (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput Biol* 6:e1000763.

[6] Whitford PW, Noel JK, Gosavi S, Schug A, Onuchic JN (2009) An All-Atom Structure-Based Potential for Proteins: Bridging Minimal Models with Empirical Forcefields. *Proteins* 75:430–441.

[7] Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN (2010) SMOG@ctbp: Simplified deployment of structure-based models in GROMACS. *Nucleic Acid Res* 38: W657-661.

[8] Lammert H, Schug A, Onuchic JN (2010) Robustness and generalization of structure based models for protein folding and function. *Proteins* 77: 881–891

[9] Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: What determines the structural details of the transition state ensemble and En-route intermediates for protein folding? An Investigation for small globular proteins. *J Mol Biol* 298: 937–953.

[10] Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* 31: 1695–1697.

[11] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comp Chem* 26: 1701-1718.

[12] Levitt M (1976) A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding. *J Mol Biol* 104: 59–107.

[13] Rotkiewicz P, Skolnik J (2008) Fast Procedure for Reconstruction of Full-Atom Protein Models from Reduced Representations *J Comp Chem* 15: 1460–1465.

[14] Hardin C, Eastwood MP, Prentiss MC, Luthey-Schulten Z, Wolynes PG (2003) Associative memory Hamiltonians for structure prediction without homology: a/b proteins. *Proc Natl Acad Sci USA* 100: 1679–1684.

[15] Eastwood MP, Hardin C, Luthey-Schulten Z, Wolynes PG (2003) Statistical mechanical refinement of protein structure prediction schemes. II. Mayer cluster expansion approach. J Chem Phys 118: 8500–8511.

[16] Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG (2004) Water in protein structure prediction. *Proc Natl Acad Sci USA* 101: 3352–3357.

[17] Kolinski A, Skolnick J (2004) Reduced models of proteins and their applications. *Polymer* 45: 511–524.

[18] Zhang Y, Kolinski A, Skolnick J (2003) Statistical mechanical refinement of protein structure prediction schemes. II. Mayer cluster expansion approach. *Biophys J* 85: 1145–1164.

[19] Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA* 98: 10125–10130.

[20] Sulkowska JI, Cieplak M (2008) Selection of optimal variants of Go-like models of proteins through studies of stretching. *Biophys J* 95: 3174–3191.

[21] Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79: 926–935.

[22] Essmann U, et al. (1995) A smooth particle Ewald method. *J Chem Phys* 103: 8577–8593.